Optimal Cache Deployment for Video-on-Demand Delivery in Optical Metro-Area Networks

Omran Ayoub*, Francesco Musumeci*, Davide Andreoletti†, Marco Mussini‡,
Massimo Tornatore* and Achille Pattavina*
*Politecnico di Milano, Department of Electronics, Information and Bioengineering, Milan, Italy
†University of Applied Sciences of Southern Switzerland, Manno, Switzerland ‡SM-Optics, Vimercate (MB), Italy

***Abstract*—Traffic demand in fixed and mobile networks is increasing rapidly, driven especially by the growing adoption of Video-on-Demand (VoD) services, which are responsible for roughly 70% of today's Internet's traffic. Network operators must continuously explore new architectural solutions to satisfy increasing traffic at minimum cost. A promising solution consists in deploying caches at the network edge such that VoD requests can be terminated locally. The dimensioning of edge network nodes in terms of storage capacity as well as their placement in the network must be optimized, to reduce costs, improve quality of service, and utilize network resources efficiently. In this paper, we aim to find the optimal deployment of caches, which minimizes overall network resource occupation for VoD service, across the various levels of a hierarchical optical metro network, in terms of the number of caches, their location and dimension (i.e., storage capacity). We develop a discrete-event simulator for dynamic VoD provisioning to measure the performance of different cache deployment strategies in terms of overall network resource occupation and blocking probability. We prove that deploying all the available storage capacity in nearest cache locations does not guarantee the minimal resource occupation. In fact, to minimize resource occupation given a fixed budget in terms of storage capacity, storage capacity must be distributed strategically among caches at different layers of the metro network based on the characteristics of the service, e.g., VoD content catalog popularity distribution.**

I. Introduction

Internet traffic keeps steadily increasing, fueled by the emergence of new bandwidth-hungry services. Among these new services, the strongest pressure on the network infrastructure is posed by VoD, which currently is responsible for 70% of Internet traffic [1]. Network operators are exploring new architectural solutions to provide users with larger capacity and improved Quality of Service (QoS) while keeping network-resource occupation low and avoiding excessive costs. A promising solution, usually referred as Network Caching, consists in enhancing nodes at the edge of the network with storage and computing capabilities [2]. Taking advantage of Network Function Virtualization (NFV) and Cloud Computing, caching enables edge network nodes to terminate services locally [3]. In this context, VoD contents are stored and streamed from *caches* deployed in the edge network nodes. Caching proved to be effective to offload traffic [4], reduce overall network energy consumption [5] [6] and as well to improve the Quality of Experience (QoE) provided to end users [7].

However, deploying a high number of large-capacity caches in edge nodes requires a very large economical investment. Consider, as an example, that a content has to be stored in its various video qualities, where certain qualities require a huge storage capacity (e.g., several GBytes per hour of 4K video even using the latest and most advanced encoding/compression algorithms). To maximize the return on the investment on cache deployment, it is decisive to choose the cache deployment strategy which allows operators to minimize resource consumption for a given investment. *With deployment strategy we refer here to choosing the number of caches, their locations (where in the network) and the size of the caches (how much storage capacity)*. An effective deployment strategy must consider network topology, users requirements and characteristics of the service (e.g., size and the popularity distribution of video content catalog).

We concentrate our analysis on metro-area networks, which are currently evolving from a rigid ring-based aggregation infrastructure to a composite cloud-network ecosystem where new services, as VoD, can be implemented and supported. Current metro networks feature several hierarchical layers (see Fig. 1, that we will describe in detail in the rest of the paper), and, when deploying caches, it is not a trivial task to decide the number of caches to be deployed, where they should be located and how to distribute the available storage capacity among the caches of the different network layers. Hence, we consider a hierarchical optical metro network hosting a VoD caching system, and our objective is to find the optimal cache deployment that minimizes network resource-occupation for a given investment in terms of cache deployment cost.

To address this problem, while taking into consideration network limitations such as link bandwidth, we develop a discrete event based simulator for dynamic VoD provisioning. The simulator generates video traffic requests based on the service characteristics (e.g., VoD content popularity model) and provisions them based on network status (e.g., available bandwidth on links) and cache deployment strategy. The simulator gives as an output the overall network resource (i.e., capacity) occupation and the blocking probability.

A. Paper Contribution and Organization

We summarize our contributions in this work as follows:

- We develop an event-based dynamic simulator for VoD content caching and distribution simulator. The simula-

tor generates video requests according to VoD content catalog popularity model and provisions them according to the implemented cache deployment strategy. The simulator also adopts adaptive bit-rate functionalities and captures the chunk nature of VoD requests.
- We demonstrate through dynamic simulations of VoD traffic that, given a budget in terms of storage capacity, the cache deployment that minimizes the overall network capacity occupation is not achieved by deploying all the available storage capacity in nearest cache locations but by deploying part of the storage capacity in caches at higher network levels.
- We also show that this optimal cache deployment improves network performance in terms of blocking probability (i.e., percentage of provisioned requests).

The rest of the paper has the following structure. Sec. II discusses some relevant works. Sec. III describes the network and VoD content catalog models. In Sec. IV we formally state the problem under investigation and in Sec.V we present the event-based simulator developed for the dynamic VoD content caching and distribution. Sec. VI reports simulative results. Sec. VII concludes the paper.

II. Related Work

Several studies addressed the problem of cache placement in telco networks. For example, Ref. [8] solves the cache deployment problem considering a trade-off between the cost of the cache deployment and that of bandwidth and energy resources. Refs. [6]-[9] address the content placement problem focusing on offloading traffic or energy-efficiency. However, these works assume either the location or the dimension of the cache is given. Ref. [10] qualitatively evaluates the impact of content caching inside telco networks in terms of cost and throughput improvement, however no budget-constraint is considered in the study. In our work, the problem is significantly different as we aim to find jointly the location and dimension of the caches at different hierarchical network levels that minimizes the overall resource consumption.

Complementary to our work, prior works have investigated real time operation of the content-delivery networks on aspects such as dynamic content management [11] and load balancing [4], [12]. All these works consider that the deployment of the caches (their locations and size) is given. In this work we focus on the cache-deployment planning phase, where caches are deployed and dimensioned with the goal of minimizing the overall network resource consumption. Similar to our work, Ref. [13] presents an optimization model whose objective is to decide where to deploy caches such as to minimize cost and maximize the bandwidth saved in a hierarchical tree network. In our work we address the problem considered in ring-based hierarchical metro networks.

III. Network and VoD Content Modeling

A. Network Model

We consider a metro network spanning over four levels, as depicted in Fig. 1, with 3 main categories of metro nodes, the

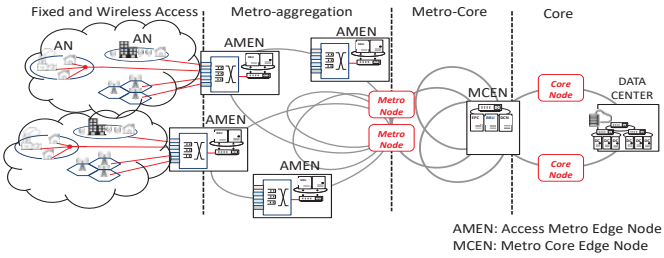

Fig. 1. Network topology considered in our study.

Access-Metro Edge Nodes (AMEN) and the Metro-Core Edge Node (MCEN) and the Metro Nodes (MN) [15]. The AMEN represents a central office where the access head-ends and the metro network interfaces are located. We assume the AMEN supports multiple access technologies and is cloud-enabled, meaning that it is equipped with storage and computing capabilities. Similarly, the MCEN is where the metro head-ends and the core network interfaces are located. It serves as a regional data center, thus it is equipped with larger cloud capabilities than those in the AMEN. In between, the MN constitute the nodes supporting the pure metro transport (no cloud capabilities). Overall, the four hierarchical levels of the network are:

- The *core* level, consisting of the core nodes and the remote data center (DC) hosting video servers. The role of the DC is to deliver contents which are not stored in caches.
- The *metro-core* level, consisting of MNs and MCENs interconnected in ring topologies.
- The *metro-access* level, consisting of AMENs and MNs connected in ring topologies.
- The *access* level, consisting of fixed and wireless access nodes, interconnected in tree topologies, where Access Nodes (ANs) represent aggregated users.

B. VoD Content Catalog Modeling

1) Catalog Size and Popularity: The catalog is characterized by its size (number of contents) and its popularity distribution. We consider a catalog size of 20.000 contents. As for the VoD content popularity (probability of a content to be requested), we assume a *Zipf*-like distribution, as [14]. The *Zipf* distribution is characterized by a small head and a long tail, meaning that a small percentage of the contents accounts for a high portion of the total requests, which motivates caching popular contents near end-users, as a small cache is sufficient to serve high amount of the VoD content requests.

2) VoD Content Characteristics: A video content is described by *i)* its popularity, *ii)* its duration and *iii)* its size. The popularity is defined as the rank of the video content in the content catalog, i.e., content #1 is the most popular content while content #20000 is the least popular content. The duration of a video content ranges between 1200 and 8400 seconds, corresponding to, e.g., a short TV-episode and a long-duration movie, respectively. The duration of a video content is assumed to follow a power-law distribution where

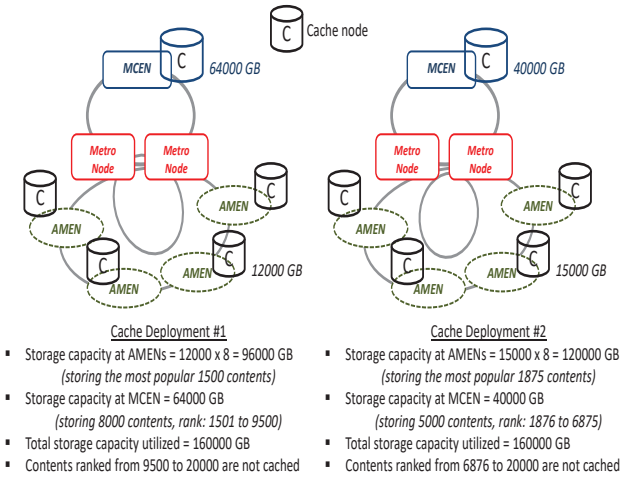

Fig. 2. Example of two cache deployments.

short videos are more common than large ones. Moreover, we assume that each video can be streamed in three different definitions, i.e., Standard Definition (SD), High Definition (HD) and Ultra High Definition (UHD), mapped to 3, 6 and 12 Mbps, respectively. We also assume a video content is stored in its best format[1], and thus its size ranges from 1.75 GB (1200 seconds x 12 Mbps) to 12.3 GB (8400 seconds x 12 Mbps) [16]. In addition, we consider the chunk-nature of a video content, where each content is made up of a number of small video-chunks. Each chunk has a fixed duration of 1.5 seconds, and the number of video-chunks in a video content can be determined by dividing its duration by the chunk duration. As an example, a video content of a duration of 1200 (or 8400) sec. consists of 800 (or 5600) chunks. As for the chunk size, we consider three different sizes (4.5, 9 or 18 Mbits) [16] according to each video quality (SD, HD or UHD).

IV. Problem Statement

The problem of optimal storage-capacity distribution in the metropolitan hierarchical network is stated as follows. **Given** a maximum overall amount of storage capacity, a metro network topology, potential location of caches and the characteristics of the content catalog (catalog size, popularity distribution), we compare different cache deployments by varying the storage capacity distribution among cache at different network layers to **find** the optimal storage capacity distribution such that the overall average Resource Occupation (*RO*) is **minimized**. We consider a dynamic VoD content distribution scenario such as to consider the limitations (e.g., link bandwidth) and the effectiveness (e.g., low blocking probability) of cache deployments. Similar to previous works (e.g., Ref. [8]), we assume the average hop-count as a main metric to estimate the overall RO, where the RO is assumed to be the product of the average hop-count and the average bit-rate.

[1]We assume all video contents are stored in one format (UHD format) and in case a lower definition is required, the content is encoded and streamed with the proper bit-rate

TABLE I
Notations considered in the analytical model.

Parameter/Variable	Description
N	Number of contents in catalog
T	Total allowed storage capacity (GB)
RO_{avg}/req	Average RO per second per VoD request
S_a	Average content size
h_a	Average hop-distance from AMEN caches
n_a	Number of AMEN caches utilized
k	Number of last content stored in AMEN caches
b_r	Average bit-rate of all VoD requests

For sake of clarity, we show in Fig. 2 two possible examples of cache deployments. We assume 160000 GB of storage capacity to be distributed among 8 caches deployed in the access network (at AMENs) and 1 cache in the metro-access network levels (at MCENs). We consider a content catalog of 20000 videos having Zipf distribution with $\alpha = 0.8$. As shown, each cache deployment results in different storing of contents in caches at AMENs and at MCEN (or left at the distribution video server). Consequently, each cache deployment results in a particular utilization of network resources. Moreover, since the number of caching nodes at different network layers varies, note that storing a content in the metro-aggregation level (i.e., at AMENs) utilizes more storage capacity with respect to storing it at the metro-core level, i.e., at MCEN (where only one copy of the content needs to be stored).

The parameters and variables considered in our study are shown in Tab. I. N represents the total number of contents in the catalog, whereas T is the total amount of storage capacity (GB) that can be deployed in caches at the various network levels. h_a, represents the average hop-distance from caches at AMENs. n_a represents the number of caches utilized in the AMENs, referred to as access caches, meaning that not necessarily all AMENs are equipped with storage capacities[2]. We call k the rank, i.e., the number of the last content stored in the access caches. Equivalently, k also represents the number of contents stored in the access caches, as the caching is popularity-based. Note that we assume the same content cannot be cached at caches of two network levels simultaneously while all caches of the same level are assumed to store the same contents. For example, if the last content stored in the access caches has rank 100 ($k = 100$), it means that all contents with rank less than 100 are stored in the access caches. Then, the number of contents which could be stored in the metro cache (given that k contents are stored in n_a access caches) is $\lfloor \frac{T}{S_a} - n_a \cdot k \rfloor$, where $\frac{T}{S_a}$ is the maximum number of contents which could be stored or duplicated and $n_a \cdot k$ is the total storage capacity (represented by number of contents) in the n_a access caches. We denote by RO_{avg}/req the average resource occupation of a video request per second under a given storage capacity distribution. RO_{avg}/req is the product of the average number of hops and the average bit-rate of all video requests.

[2]For example, in a topology of 32 AMENs, it is possible that only half of them are equipped with storage capacity thus having $n_a = 16$. In such a case, caches at AMENs are placed such as the maximum number of hops between an end-user and AMEN cache is minimized

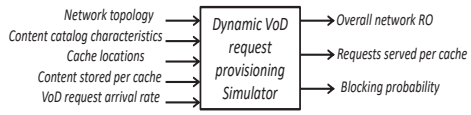

Fig. 3. The overall framework of the discrete-event based simulator for VoD content caching and distribution.

V. Dynamic VoD Content Caching and Distribution Simulator

To identify the cache deployment that minimizes the overall network capacity occupation, we perform dynamic simulations of VoD traffic requests over a network architecture as in Sec. III. In this section we present a description of the discrete-event-driven simulator for dynamic VoD content caching and distribution, and we explain the VoD request provisioning process.

The overall framework of the simulator is represented in Fig. 3. Given the network topology, content catalog characteristics, locations of caches and the list of stored contents, the simulator provisions the dynamically-arriving VoD requests, based on current network status, and gives as an output overall RO, requests served by each cache and blocking probability.

The physical topology is described by a graph G=(V,E), where V is the set of nodes and E the set of edges, representing network nodes and physical links, respectively. The VoD content request is described by the tuple r = (t_s, D_r, m, b_r, d_r), where t_s is the request arriving time of user D_r, m is the content requested with bit-rate b_r having a chunk duration of d_r. Note that a content consists of a number of chunks and the chunks are provisioned one by one. Moreover, we note that according to the cache deployment strategy, different caches might be storing the requested content. Consequently, an anycast routing problem arises. To transform the anycast routing into a unicast routing, we introduce in our work an anycast abstraction of the topology [17].

The simulated VoD-chunk provisioning/deprovisioning process is shown in Alg. 1. When a user requests a content m, a list of all cache nodes hosting m (including the video server) is identified. Then, the nearest cache storing the content delivers the content to user D_r, considering a path with available bandwidth greater than or equal to b_r (line 5). If no path with bandwidth greater than or equal to b_r is found, the provisioning process tries to accommodate the request degrading the quality of the video delivery, i.e., setting the value of b_r to that corresponding to the lower video definition (described in section III-B2). If a path is found, request r is provisioned for the duration of the chunk of the content and then it is deprovisioned at time $t_s + d_r$ deallocating the assigned bandwidth from the utilized path. Simultaneously, a request for the successive chunk of the content request is initiated. However, if the request was not provisioned, i.e., no path with enough bandwidth is available, the VoD request is blocked (lines 20-21). Note that since the provisioning/deprovisioning process is performed for every chunk of a VoD request, different chunks can be delivered at different provisioning bit-rates, thus imitating the functionality of the adaptive bit-rate streaming technique. For example, if a chunk is delivered with a bit-rate lower than b_r, due to unavailable network resources, the successive chunk shall be allocated bit-rate b_r if network resources became available.

Algorithm 1 VoD-Chunk Provisioning

Input: *Network status: caches location, stored contents per cache and available bandwidth on links. VoD content request: $r(t_s, D_r, m, b_r, d_r)$: arriving time t_s, user D_r, requested content m, requested bit-rate b_r, chunk duration d_r.*
Output: *VoD provisioning (per chunk)*

1: *provision* = $false$;
2: *assigned bandwidth* = 0;
3: Allocate a list of caches storing content m;
4: *path is found* = $false$;
5: Find shortest path between D_r and nearest cache node with available bandwidth $\geq b_r$;
6: **if** (shortest path found with available bandwidth $\geq b_r$) **then**
7: *path is found* = $true$;
8: *assigned bandwidth* = b_r;
9: **else**
10: Find shortest path between D_r and cache node with available bandwidth $\geq min.\ bandwidth$;
11: **if** (shortest path found with available bandwidth $\geq min.\ bandwidth$) **then**
12: *path is found* = $true$;
13: *assigned bandwidth* = $min.\ bandwidth$;
14: **end if**
15: **end if**
16: **if** (*path is found* = $true$) **then**
17: Provision r over the shortest path assigning bandwidth *assigned bandwidth*;
18: Automatically initiate request for the next video-chunk at time $t_s + d_r$;
19: Deprovision r at time $t_s + d_r$, i.e., release bandwidth *assigned bandwidth* from considered path;
20: **else**
21: Block request r;
22: **end if**
23: End;

VI. Numerical Results

A. Optimal Storage-Capacity Distribution

In this section we aim to find the optimal storage-capacity distribution for three different case studies (whose parameters are reported in Tab. II), which differ in the available storage capacity (the budget of the cache deployment) and the service characteristics (i.e., different popularity model and content catalog size by varying α and S_a). The considered network topology (similar to the one depicted in Fig. 1) consists of 1 MCEN, 4 metro nodes, 32 AMENs and 96 ANs distributed over 4 different access rings. VoD requests are originated uniformly from all ANs with probability 0.5, 0.25 and 0.25 of choosing bit-rates of 3 Mbps (SD), 6 Mbps (HD) or 12 (Ultra HD) Mbps, respectively. Consequently, the average bit-rate of all requests is 6 Mbps. Network links support 2·10 Gigabit Ethernet (GE) technology in the *core* and the *metro* transport links and 10 GE is adopted in the *access-metro* ring. Another

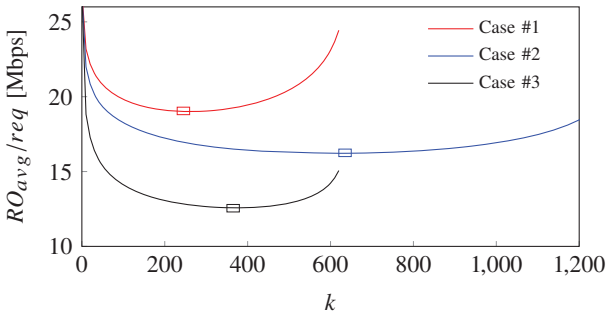

Fig. 4. RO_{avg}/req with respect to the number of contents stored in the AMEN caches for the three case studies.

important input parameter which we vary is the number caches located at AMENs (n_a). We consider that either 16 or 32 AMENs host caches. Note that if all AMENs are equipped with caches ($n_a = 32$), the average hop-distance from the AMEN caches $h_a = 1$, whereas if half of the AMENs are equipped with caches, ($n_a = 16$), $h_a = 1.5$.

We perform different simulations for different cache deployments. In the simulations we vary the number of contents stored in caches deployed at AMENs, k. k ranges from 0 (the case where all the storage capacity is utilized in the cache located at the MCEN), to $\frac{T}{n_a \cdot S_a}$, the maximum number of contents to store in each of the AMEN caches, i.e., the case where all the storage capacity is utilized in the caches located at the AMENs. We refer to these cache deployments by *Only MCEN* and *Only AMENs*, respectively. Note that when a certain amount of available storage capacity is utilized in the caches of the AMENs, the remaining amount is utilized in the cache located at the MCEN. In each simulation, we simulate the arrival of 400000 VoD requests, assumed as Poisson-distributed, at an arrival rate guaranteeing negligible blocking probability, so as to provide a fair comparative analysis between the different cache deployments.

Figure 4 shows RO_{avg}/req (i.e., the average resource occupation per VoD request) as a function of the number of contents stored in the caches hosted by AMENs, k, for all case studies in Tab. II. Results, in all case studies, show that RO_{avg}/req initially decreases as k increases (as more contents stored in the caches located at AMENs allow to serve more requests from locations near end-users) until a certain value of k, about which RO increases again (as it becomes less-advantageous to deploy more storage capacity in the caches of the AMENs and more-advantageous to deploy the storage capacity in the MCEN cache). Why is the optimal solution is not deploying storage capacity at AMENs? It is due to the fact that, when the storage capacity is limited, it becomes more-advantageous not to store duplicates of a number of popular contents at AMENs but rather store one

TABLE II
SIMULATION SETTINGS FOR THE CONSIDERED CASE STUDIES.

Case Study #	α	S_a (GB)	T	$AMENs$	n_a	h_a
1	0.9	8	160000	32	32	1
2	0.9	6	120000	32	16	1.5
3	1	8	160000	32	32	1

TABLE III
VALUES OF k_* AND STORAGE CAPACITY OF AMENS AND MCEN CACHES FOR EACH CASE STUDY.

Case Study #	k_*	AMENs Cache (GB)	MCEN Cache (GB)
1	267	2136	91648
2	615	3690	60960
3	377	3016	63488

copy of a larger set of contents, thus pulling more contents from the origin server into the network.

In Tab. III we show the value of k^*, i.e., the number of contents that, if stored in the AMEN caches, guarantees an optimized cache deployment, and the storage capacity of the caches located at AMENs and MCEN for each of the considered case study. First, we highlight through comparing the cache deployment for cases #1 and #2 that any variation in the characteristics of the content catalog (in this case the average content size S_a) or the storage-capacity budget (available amount of storage capacity T) changes the optimal solution for cache deployment. For example, in case #1 a relatively low number of contents is stored in AMEN caches (and a large MCEN cache is preferred) while in case #2 more contents are stored in AMENs (and a smaller MCEN cache is preferred). Comparing cases #1 and #3, which differ only in the VoD content catalog popularity distribution (i.e., skew parameter α), we can draw another conclusion: the popularity distribution of the VoD content catalog drastically changes the optimal cache deployment as the deployment of a larger AMEN caches (and a smaller MCEN cache) is preferred in case #3 with respect to case #1.

B. Optimized Cache Deployment vs. Baseline Strategies

Now we evaluate the advantage of having an optimized cache deployment, referred to as *k-optimized*, through comparing it to two baseline strategies (*Only MCEN* and *Only AMENs*) in terms of the blocking probability (*Pb*). Considering case study #1, we simulate the arrival of 400000 Poisson-distributed VoD requests with an arrival rate varying from 10 to 100 requests per second[3]. Fig. 5 shows the *Pb* of the cache deployments with respect to the carried load. We notice that *Only MCEN* has the worst performance due to the fact that, since only one cache is deployed at metro level, the traffic overloads the links connected to the MCEN cache. *Only AMENs* shows a better performance with respect to *Only MCEN* yet not as good as the *k-optimized* deployment which outperforms the other deployments, as it requires less network resources with respect to other deployments.

Tab. IV shows the percentage of requests served from the data center (DC), the MCEN cache, and the AMENs caches for the three cases in Tab. II under different cache deployments strategies, namely; the *Only MCEN*, *Only AMENs* and *k-optimized*. In all cases, for *Only MCEN* cache deployment, all the contents are pulled from the DC in the MCEN, as the available storage capacity allows storing and serving all

[3]Note that the Poisson distribution here refers only to the arrival of new requests while the the chunks of an already provisioned request arrive after a deterministic interval equal to the chunk's duration.

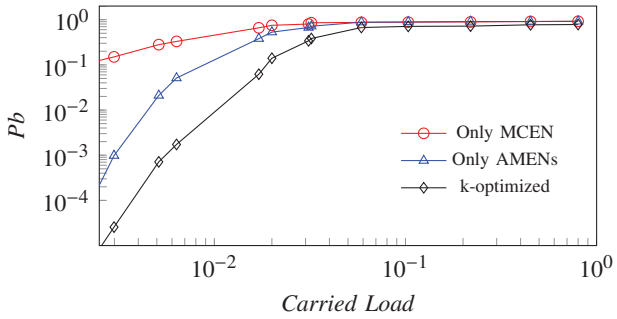

Fig. 5. *Pb* with respect to the carried load for different cache-deployment strategies for case study #1.

TABLE IV
PERCENTAGE OF REQUESTS SERVED FROM THE DC, THE MCEN CACHE AND THE AMENS CACHES AND AVERAGE RO FOR EACH CACHE DEPLOYMENT IN THE THREE CASE STUDIES.

Case Study #1				
Cache Deployment	DC	MCEN Cache	AMENs Caches	RO
Only MCEN	0	100%	0	27
$k-optimized$	9%	46%	45%	19
Only AMENs	42%	0	58%	24.43
Case Study #2				
Cache Deployment	DC	MCEN Cache	AMENs Caches	RO
Only MCEN	0	100%	0	27
$k-optimized$	4%	27%	69%	16.21
Only AMENs	27%	0	73%	19.32
Case Study #3				
Cache Deployment	DC	MCEN Cache	AMENs Caches	RO
Only MCEN	0	100%	0	27
$k-optimized$	5%	22%	73%	12.57
Only AMENs	22%	0	78%	15.06

the contents from the MCEN cache. This cache deployment results in the highest overall network *RO*. For the *Only AMENs*, the storage capacity is divided among the AMEN caches, storing the most popular contents. Although this cache deployment allows serving end-users a high percentage of the requests (ranging between 58% and 78%) from close locations, it leaves a significant percentage of requests (ranging between 22% and 42%) to be served from the DC, consequently increasing the overall network *RO*. As for the *k-optimized* cache deployment, we see that it reveals a lower percentage of requests served from the AMEN caches with respect to that of the *Only AMENs* cache deployment but a much higher percentage of requests served from the MCEN cache. Consequently, only a low percentage of requests is served from the DC, thus resulting in the minimal overall network *RO*.

VII. CONCLUSION

In this paper we address the problem of finding the optimal cache deployment in terms of number of caches, their location and through distributing the available capacity storage across the caches of the hierarchical levels of an optical metro network such that the overall network resource occupation is minimized. To this end we developed a discrete-event simulator for dynamic VoD content caching and distribution and performed dynamic simulations of different cache deployments. Simulative results show that deploying the available storage capacity only in the access-aggregation segment (i.e., closer to end-users) does not guarantee a minimized overall network resource occupation. This is because such a cache deployment maximizes the hit-ratio of a significant number of caches in the access segment, consuming most of the available storage capacity and leaving a huge percentage of video contents at the origin data center. In fact, results show that distributing the available storage capacity among access and metro caches following the optimal cache deployment yields a better resource occupation and a lower blocking probability. Results also quantitatively show the effect of characteristics of the service (e.g., VoD content catalog popularity distribution) on the optimal cache deployment. As future work, we aim to formulate and solve the cache deployment problem in hierarchical metro networks analytically.

ACKNOWLEDGMENT

The work leading to these results has been supported by the European Community under grant agreement no. 761727 *Metro-Haul* project and the *Lombardy region* through *New Optical Horizon* project funding.

REFERENCES

[1] Webster, D. (2017). Cisco Visual Networking Index (VNI). Global Forecast Update, 6.
[2] Peterson, Larry, et al. "Central office re-architectured as a data center." IEEE Communications Magazine, 96-101, 2016.
[3] Tao, M., et al. Communications, caching, and computing for content-centric mobile networks: part 1. IEEE Communications Magazine, 14-15, 2016.
[4] Sourlas, V., et al. Distributed cache management in information-centric networks. IEEE Transactions on Network and Service Management, 10(3), 286-299, 2013.
[5] Mathew, V., et al. "Energy-aware load balancing in content delivery networks." In INFOCOM, Proceedings IEEE (pp. 954-962), 2012.
[6] Savi, M., et al. "Energy-efficient caching for Video-on-Demand in Fixed-Mobile Convergent networks." In Green Communications (OnlineGreenComm), IEEE Online Conference on, 2015.
[7] Yang, C., et al. "A game theoretical framework for improving the quality of service in cooperative RAN caching." In Communications (ICC),International Conference on, IEEE, 2017.
[8] Hasan, S., et al. "Trade-offs in optimizing the cache deployments of CDNs." In INFOCOM, Proceedings IEEE, 2014.
[9] Llorca, J., et al. "Dynamic in-network caching for energy efficient content delivery." In INFOCOM, 2013 Proceedings IEEE, 2013.
[10] Ciccarella, Gianfranco et al. "Performance improvement and network TCO reduction by optimal deployment of caching." Euro Med Telco Conference (EMTC). IEEE, 2014.
[11] Applegate, D., et al. "Optimal content placement for a large-scale VoD system." In Proceedings of the 6th International Conference. ACM, 2010.
[12] Sourlas, V., et al. "Distributed cache management in information-centric networks." IEEE Transactions on Network and Service Management, 10(3), 286-299, 2013.
[13] Gourdin, Eric et al. "Optimal hierarchical deployment of caches for video streaming." Network of the Future (NOF), 6th International Conference on the. IEEE, 2015.
[14] Li, H., et al. "Video requests from online social networks: Characterization, analysis and generation." In INFOCOM, Proceedings (pp. 50-54). IEEE, 2013.
[15] Metro-HAUL project, https://metro-haul.eu/
[16] Casas, P., et al. "When YouTube does not work: Analysis of QoE-relevant degradation in Google CDN traffic." IEEE Transactions on Network and Service Management, 11(4), 441-457, 2014.
[17] Gattulli, Mirko, et al. "Low-emissions routing for cloud computing in IP-over-WDM networks with data centers." IEEE Journal on Selected Areas in Communications, 28-38, 2014.